Semantic Activity Recognition

Monique Thonnat

▶ **To cite this version:**

Monique Thonnat. Semantic Activity Recognition. 18th European Conference on Artificial Intelligence ECAI 2008, Jul 2008, Patras, Greece. pp 3-7. ⟨inria-00502249⟩

HAL Id: inria-00502249

https://inria.hal.science/inria-00502249v1

Submitted on 13 Jul 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

Semantic Activity Recognition

Monique Thonnat [1]

Abstract.

Extracting automatically the semantics from visual data is a real challenge. We describe in this paper how recent work in cognitive vision leads to significative results in activity recognition for visualsurveillance and video monitoring. In particular we present work performed in the domain of video understanding in our PULSAR team at INRIA in Sophia Antipolis. Our main objective is to analyse in real-time video streams captured by static video cameras and to recognize their semantic content. We present a cognitive vision approach mixing 4D computer vision techniques and activity recognition based on a priori knowledge. Applications in visualsurveillance and healthcare monitoring are shown. We conclude by current issues in cognitive vision for activity recognition.

1 INTRODUCTION

This paper is focused on activity recognition. Activity recognition is a hot topic in the academic field not only due to scientific motivations but also due to strong demands coming from the industry and the society; in particular for videosurveillance and healthcare. In fact, there is an increasing need to automate the recognition of activities observed by visual sensors (usually CCD cameras, omni directional cameras, infrared cameras). More precisely we are interested in the **real-time semantic interpretation of dynamic scenes** observed by video cameras. We thus study spatio-temporal activities performed by mobile objects (e.g. human beings, animals or vehicles) interacting with the physical world.

What does it mean to understand a video ? Is it just to perform statistics on the appearance of images and to recognize an image from a set of already seen images? If we really want to understand the activities performed by the physical objects 2D analysis is not sufficient. We need to locate the physical objects in the 3D real world. The dynamics of the physical objects is a major cue for activity recognition. The computer vision community is very active in the domain of motion detection, mobile object tracking and more recently trajectory analysis. Very often these analyses are performed in the image plane and are thus dependant of the sensor parameters as its field of view, position and orientation. However for reliable activity recognition the dynamics of the physical objects must be computed in the 4D space.

Is there a unique objective interpretation of a dynamic scene? For instance the scenes shown in figures 1 and 2 can be interpreted more or less precisely in function of the a priori knowledge of the observer. In the first case (shown in figure 1) without information on the location of the scene one can recognize an indoor scene where two men are walking together towards a door; a videosurveillance expert knowing the location (a bank agency), its spatial configuration as well as security rules will interpret the same scene as a bank attack with the unautorized person accessing together with an employee to a fordidden area. In the second case (shown in figure 2) without information on the location of the scene one can recognize a woman standing alone; a medical expert knowing the patient will interpret the same scene as an active elderly preparing a meal in her kitchen. In fact, the interpretation of a video sequence is not unique but it depends on the a priori knowledge of the observer and on his/her goal.

Figure 1. A scene with different valid interpretations: two people walking together towards a door or a bank attack with an access to a forbidden area by an unauthorized person and an employee.

Figure 2. A scene with different valid interpretations: a person standing in a room or an active elderly preparing a meal in a kitchen.

2 4D APPROACH

We present a cognitive vision approach mixing 4D computer vision techniques and activity recognition based on a priori knowledge. The major issue in semantic interpretation of dynamic scenes is the gap between the subjective interpretation of data and the objective measures provided by the sensors.

[1] INRIA, France, email: Monique.Thonnat@sophia.inria.fr

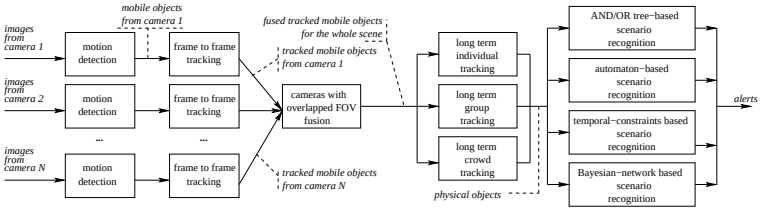
images from camera 1
images from camera 2
images from camera N
motion detection
frame to frame tracking
mobile objects from camera 1
tracked mobile objects from camera 1
tracked mobile objects from camera N
fused tracked mobile objects for the whole scene
cameras with overlapped FOV fusion
long term individual tracking
long term group tracking
long term crowd tracking
physical objects
AND/OR tree−based scenario recognition
automaton−based scenario recognition
temporal−constraints based scenario recognition
Bayesian−network based scenario recognition
alerts

Figure 3. From sensor data to high level interpretation; global structure of an activity monitoring system built with VSIP[1].

Our approach to address this problem is to keep a clear boundary between the application dependent subjective interpretations and the objective analysis of the videos. We thus define a set of objective measures which can be extracted in real-time from the videos, we propose formal models to enable users to express their activities of interest and we build matching techniques to bridge the gap between the objective measures and the activity models.

Figure 3 shows the global structure of a videosurveillance system built with this approach. First, a motion detection step followed by a frame to frame tracking is made for each video camera. Then the tracked mobile objects coming from different video cameras with overlapping fields of view are fused into a unique 4D representation for the whole scene. Depending on the chosen application, a combination of one or more of the available trackers (individuals, groups and crowd tracker) is used. Then scenario recognition is performed by a combination of one or more of the available recognition algorithms (automaton based, Bayesian-network based, AND/OR tree based and temporal constraints based). Finally the system generates the alerts corresponding to the predefined recognized scenarios.

For robust semantic interpretation of mobile object behaviour it is mandatory to rely on correct physical object type classification. It can be based on simple 3D models like parallelepipeds [12] or complex 3D human body configurations with posture models as in [2]. Figure 4 shows examples of such postures.

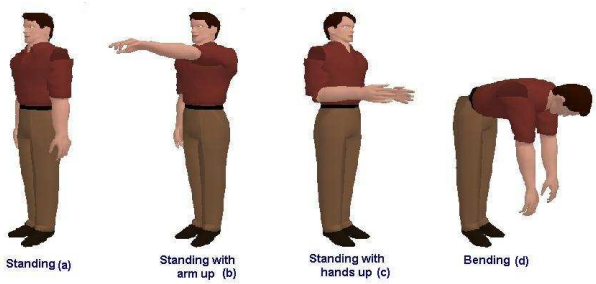

Figure 4. Different 3D models of human body postures

3 3D MAP

We use 3D maps as a means to model the a priori knowledge of the physical environment captured by the sensors. More precisely the 3D maps contain the a priori knowledge of the empty scenes:

- **Video Cameras:** 3D position of the sensors, calibration matrix, fields of view,...
- **3D Geometry:** the geometry of the static structure of the empty scene (for instance the buildings and road structure for outdoor scenes or the walls and doors for indoor scenes) as well as the main static 3D objects (for instance the furniture in indoor scenes) and the 2D zones of interest. This geometry is defined in terms of 3D position, shape and volume.
- **Semantic information:** for each part of the map semantic information is added as its type (e.g. 3D object, 2D zone), its characterics (e.g. yellow, fragile) or its function (e.g. entrance zone, seat).

We can see on figure 5 a 2D map of an indoor flat and on figure 10 two partial views of the 3D map built for monitoring elderly at home. In this map in addition to the main structure of the rooms (walls, doors, etc.), the equipment and the furniture are defined as well as the information related to the sensors.

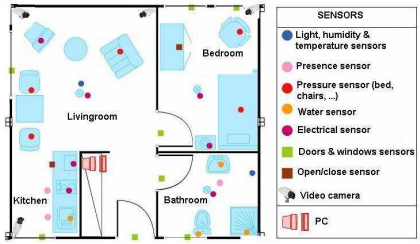

Figure 5. Top view of the flat

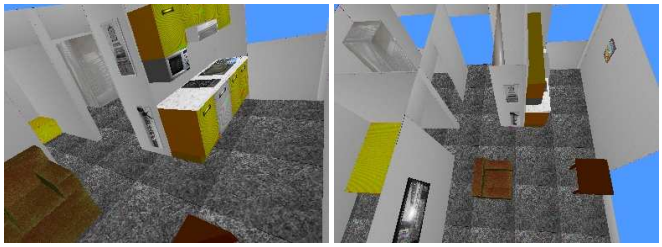

Figure 6. 3D map: the kitchen area and the top view of a flat for monitoring elderly at home

4 ACTIVITY MODELLING

In order to express the semantics of the activities a modelling effort is needed. The models correspond to the modeling of all the knowledge needed by the system to recognize video events occurring in the scene. To allow security operators to easily define and modify their models, the description of the knowledge is declarative and intuitive

(in natural terms). We propose a video event ontology to share common concepts in video understanding and to decrease the effort of knowledge modelling.

4.1 The Video Event Ontology

The event ontology is a set of concepts for describing physical objects, events and relations between concepts:

The **physical objects** are all the concepts to describe objects of the real world in the scene observed by the sensors. The attributes of a physical object are pertinent for the recognition. These attributes characterize the physical object. There are two types of physical objects: contextual objects (which are usually static and whenever in motion, its movement can be predicted using contextual information) and mobile objects (which can be perceived as moving in the scene and as initiating their motions, without the possibility to predict their movement).

The **events** are all the concepts to describe mobile object evolutions and interactions in a scene. Different terms are used to describe these concepts and categorized into two categories: **state** (including primitive/composite state) and event (including primitive/composite event, single/multi-agent event).

A **primitive state** is a spatio-temporal property valid at a given instant or stable on a time interval which is directly inferred from audiovisual attributes of physical objects computed by low level signal processing algorithms.

A **composite state** is a combination of states. A **primitive event** is a change of states. A **composite event** is a combination of states and events. A single-agent event is an event involving a single mobile object. A multi-agent event is a composite event involving several (at least two) mobile objects with different motions.

Currently this ontology contains 151 concepts used for different applications in video understanding. This ontology is implemented in Protege to be independant of a particular activity recognition formalism.

4.2 Activity Models

A formalism for expressing an activity is directly based on the concepts of the video event ontology. A composite event model is composed of five parts: "**physical objects**" involved in the event (e.g. person, equipment, zones of interest), "**components**" corresponding to the sub-events composing the event, "**forbidden components**" corresponding to the events which should not occur during the main event, "**constraints**" are conditions between the physical objects and/or the components (including symbolic, logical, spatial and temporal constraints including Allen interval algebra operators, and "**alarms**" describing the actions to be taken when the event is recognized.

Primitive states, composite states and primitive events can be described using the same formalism. Please see [10] and [9] for more details of the formalism.

5 ACTIVITY RECOGNITION

The algorithm proposed in [9] and in [10] enables to process efficiently (i.e. in realtime) a data flow and to recognize pre-defined activities. Alternative approaches based on probabilistic methods [6] or [7] can also be used. In the following we concentrate on the first approach because it is directly based on the formalism and the ontology presented in the previous section. The video event recognition algorithm recognizes which events are occurring using the primitive video events. To recognize an event composed of sub-events, given the event model, the recognition algorithm selects a set of physical objects matching the remaining physical object variables of the event model. The algorithm then looks back in the past for any previously recognized state/event that matches the first component of the event model. If these two recognized components verify the event model constraints (e.g. temporal constraints), the event is said to be recognized. In order to facilitate complex event recognition, after each event recognition, event templates are generated for all composite events, the last component of which corresponds to this recognized event. For more details see [9].

6 APPLICATIONS

This approach has been applied to a large set of applications in visualsurveillance.

6.1 Visualsurveillance

A typical example of complex activities in which we are interested is aircraft monitoring (see figure 7 in apron areas . In this example the duration of the servicing activities8 around the aircraft is about one hour and the activities involve interactions between several ground vehicles and human operators.

The goal is to recognize these activities through formal activity models as shown in figure 9 and data captured by a network of video cameras (such as the ones shown in figure 7). For more details, refer to [3] and the related European project website http://www.avitrack.net/.

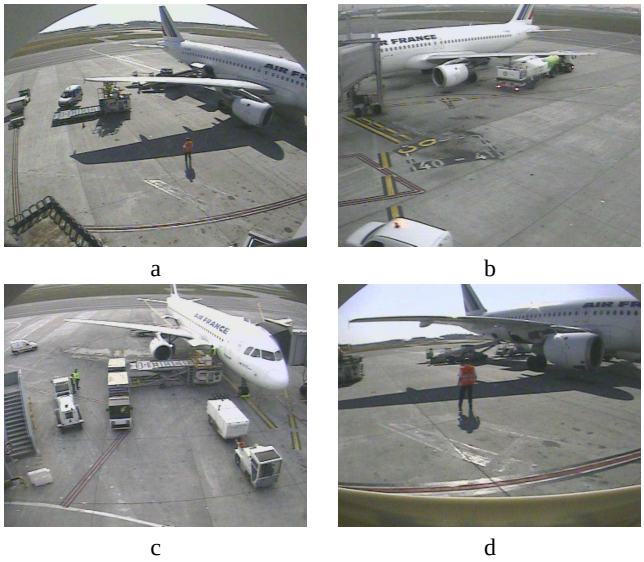

Figure 7. Different views of an apron area captured by video cameras for aircraft monitoring

6.2 Healtcare monitoring

In this application the objective is to monitor elderly at home (see figure 10). In collaboration with gerontologists, we have modeled several primitive states, primitive events and composite events. First we

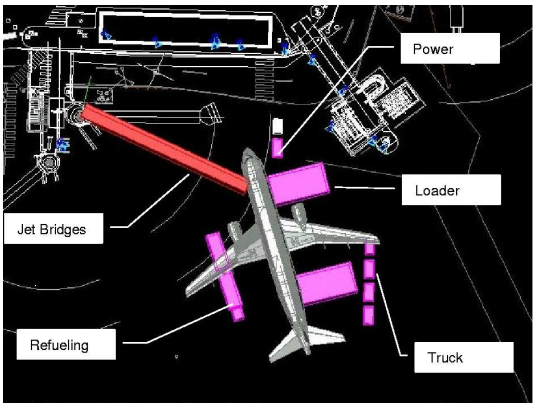

Figure 8. Activity recognition problem in airport: the main servicing operations around an aircraft (refuelling, baggage loading, power supply, etc...) and the location of the 8 video cameras (in blue)

```
CompositeEvent(Unloading_Operation,
  PhysicalObjects( (p1 : Person), (v1 : Vehicle), (v2 : Vehicle), (v3 : Vehicle),
                   (z1 : Zone), (z2 : Zone),(z3 :Zone), (z4 : Zone))
  Components(  (c1 : CompositeEvent Loader_Arrival(v1, z1, z2))
               (c2 : CompositeEvent Transporter_Arrival(v2, z1, z3))
               (c3 : CompositeState  Worker_Manipulating_Container(p1, v3, v2, z3, z4)))
  Constraints(  (v1->SubType = LOADER)
                (v2->SubType = TRANSPORTER)
                (z1->Name = ERA)
                (z2->Name = Front_Loading_Area)
                (z3->Name = Transporter_Area)
                (z4->Name = Container_Worker_Area)
                (c1 before_meet c2)
                (c2 before_meet c3)))
```

Figure 9. Activity recognition problem in airport: example of an activity model enabling to describe an unloading operation with a high-level language

Figure 10. healthcare

are interesting in modelling events characteristic of critical situations such as falling down. Second, these events aim at detecting abnormal changes of behavior patterns such as depression. Given these objectives we have selected the activities that can be detected using video cameras [11]. We have modeled **thirty four video events**. In particular, we have defined fourteen primitives states, four of them are related to the location of the person in the scene (e.g. inside kitchen, inside livingroom) and the ten remaining are related to the proposed 3D key human postures. We have defined also four primitive events related to the combination of these primitive states: **"standing up"** which represents a change state from sitting or slumping to standing, **"sitting down"** which represents a change state from standing, or bending to sitting on a chair, **"sitting up"** represents a change state from lying to sitting on the floor, and **"lying down"** which represents a change state from standing or sitting on the floor to lying. We have defined also six primitive events such as: stay in kitchen, stay in livingroom. These primitive states and events are used to define more composite events. For this study, we have modeled ten composite events. In this paper, we present just two of them: **"feeling faint"** and **"falling down"**.

The model of the "feeling faint" event is shown in figure 4. The "feeling faint" model involves one physical object (one person), and it contains three 3D human posture components and constraints between these components.

CompositeEvent (PersonFeelingFaint,
PhysicalObjects((p: Person))
Components
((pStand: PrimitiveState Standing(p))
(pBend: PrimitiveState Bending(p))
(pSit: PrimitiveState Sitting_Outstretched_Legs(p)))
Constraints
((Sequence pStand; pBend; pSit)
(pSit's Duration >= 10))
Alarm(AText("Person is Feeling Faint")
AType("URGENT")))

"Feeling faint" model.

We have also modelled the "falling down" event. There are different ways for describing a person falling down. Thus, we have modelled the event "falling down" with three models:

Falling down 1: A change state from standing, sitting on the floor (with flexed or outstretched legs) and lying (with flexed or outstretched legs).

Falling down 2: A change state from standing, and lying (with flexed or outstretched legs).

Falling down 3: A change state from standing, bending and lying (with flexed or outstretched legs).

An example of the definition of the model "falling down 1" is shown below.

CompositeEvent(PersonFallingDown1,
PhysicalObjects((p: Person))
Components
((pStand: PrimitiveState Standing(p))
(pSit: PrimitiveState Sitting_Flexed_Legs(p))
(pLay: PrimitiveState Lying_Outstretched_Legs(p)))
Constraints
((pSit before_meet p_lay)
(pLay's Duration >= 50))
Alarm
(AText("Person is Falling Down")
AType("VERYURGENT")))

"Falling down 1" model.

Figure 11 and figure 12 show respectively the camera view and the 3D visualization of the recognition of the "feeling faint" event.

Figure 11. Recognition of the ”feeling faint” event

Figure 12. 3D visualization of the recognition of the ”feeling faint” event

Figure 13 and figure 14 show respectively the camera view and the 3D visualization of the recognition of the ”falling down” event.

Figure 13. Recognition of the ”falling down” event

Figure 14. 3D visualization of the recognition of the ”falling down” event

7 CONCLUSION

We have shown a 4D semantic approach for activity recognition of dynamic scene. There are still a lot of open issues among which a full theory of visual data interpretation, reliable techniques for 4D analysis able to deal with changing observation conditions and scene content. From an activity recognition point of view the three main points are the development of shared operational ontologies, of formalisms for activity modelling with good properties such as scalability and learning techniques for model refinement. In particular a large set of learning issues are rised by this 4D semantic approach for instance: learning contextual variations for physical object detection and image segmentation [5], learning the structure of the activity models [8] or learning the visual concept detectors [4].

REFERENCES

[1] A. Avanzi, F. Bremond, C. Tornieri, and M. Thonnat, ‘Design and assessment of an intelligent activity monitoring platform’, *EURASIP Journal on Applied Signal Processing, special issue in ”Advances in Intelligent Vision Systems: Methods and Applications”*, **2005**(14), 2359–2374, (August 2005).

[2] B. Boulay, F. Bremond, and M. Thonnat, ‘Applying 3d human model in a posture recognition system’, *Pattern Recognition Letter, Special Issue on vision for Crime Detection and Prevention*, **27**(15), 1788–1796, (2006).

[3] Florent Fusier, Valery Valentin, François Bremond, Monique Thonnat, Mark Bor g, David Thirde, and James Ferryman, ‘Video understanding for complex activity recognition’, *Machine Vision and Applications Journal*, **18**, 167–188, (2007).

[4] N. Maillot and M. Thonnat, ‘Ontology based complex object recognition’, *Image and Vision Computing Journal, Special Issue on Cognitive Computer Vision*, **26**(1), 102–113, (2008).

[5] V. Martin and M. Thonnat, ‘Learning contextual variations for video segmentation’, in *The 6th International Conference on Vision Systems (ICVW08)*, Santorini, Greece, (2008).

[6] G. Medioni, I. Cohen, F. Brémond, S. Hongeng, and G. Nevatia, ‘Activity Analysis in Video’, *Pattern Analysis and Machine Intelligence PAMI*, **23**(8), 873–889, (2001).

[7] N. Moenne-Loccoz, F. Brémond, and M. Thonnat, ‘Recurrent bayesian network for the recognation of human behaviors from video’, in *Third International Conference On Computer Vision Systems (ICVS 2003)*, volume LNCS 2626, pp. 44–53, Graz, Austria, (2003). Springer.

[8] A. Toshev, F. Brémond, and M. Thonnat, ‘An a priori-based method for frequent composite event discovery in videos’, in *Proceedings of 2006 IEEE International Conference on Computer Vision Systems*, New York USA, (January 2006).

[9] V-T. Vu, F. Brémond, and M. Thonnat, ‘Automatic video interpretation: A novel algorithm for temporal scenario recognition’, in *The Eighteenth International Joint Conference on Artificial Intelligence (IJCAI’03)*, Acapulco, Mexico, (2003).

[10] V-T. Vu, F. Brémond, and M. Thonnat, ‘Automatic video interpretation: A recognition algorithm for temporal scenarios based on pre-compiled scenario models’, in *The 3rd International Conference on Vision System (ICVS’03)*, Graz, Austria, (2003).

[11] N. Zouba, B. Boulay, F. Brémond, and M. Thonnat, ‘Monitoring activities of daily living (adls) of elderly based on 3d key human postures’, in *The 4th International Cognitive Vision Workshop (ICVW08)*, Santorini, Greece, (2008).

[12] M. Zúniga, F. Brémond, and M. Thonnat, ‘Fast and reliable object classification in video based on a 3d generic model’, in *The 3rd International Conference on Visual Information Engineering (VIE2006)*, pp. 433–441, Bangalore, India, (September 26-28 2006).